ANALISIS REKOMENDASI PENERIMA BEASISWA MENGGUNAKAN ALGORITMA *K-NEAREST NEIGHBOR* (*K*-NN) DAN ALGORITMA C4.5

[1]Dita Noviana, [2]Yuliana Susanti, [3]Irwan Susanto
[1,2,3]Universitas Sebelas Maret, Surakarta, (0271) 663375
e-mail: ditanoviana97@gmail.com

Abstrak

Beasiswa merupakan bantuan biaya pendidikan yang sangat membantu prestasi mahasiswa. Beasiswa yang diberikan antara lain beasiswa Peningkatan Prestasi Akademik (PPA) yang diberikan kepada mahasiswa berprestasi. Seiring dengan meningkatnya jumlah mahasiswa yang mengajukan beasiswa, maka dibutuhkan suatu metode klasifikasi yang dapat membantu menentukan siapa yang layak mendapatkan beasiswa PPA dari Universitas Sebelas Maret. Metode yang digunakan dalam analisis ini adalah *K-Nearest Neighbor* dan C4.5. Hasil klasifikasi digunakan sebagai keputusan dalam rekomendasi penerima beasiswa. Penelitian ini mengambil sampel sebanyak 1000 calon penerima beasiswa PPA Universitas Sebelas Maret, sebanyak 907 mahasiswa diklasifikasikan secara benar melalui perhitungan algoritma *k*-NN dan sebanyak 883 mahasiswa diklasifikasikan secara benar melalui perhitungan C4.5. Hasil penelitian juga menunjukkan akurasi dan eror algoritma *k*-NN sebesar 90.7% dan 9.3%, sedangkan pada algoritma C4.5 memiliki akurasi dan eror sebesar 88.3% dan 11.7%.

Kata Kunci: Klasifikasi, *K-Nearest Neighbor*, C4.5, beasiswa

Abstract

Scholarships are tuition fees that greatly help student achievement. Scholarships provided include Academic Achievement Improvement scholarships given to outstanding students. Along with the increasing number of students applying for scholarships, a classification method is needed that can help determine who is eligible for a PPA scholarship from Sebelas Maret University. The methods used in this analysis are the *K-Nearest Neighbor* and C4.5. The classification results are used as decisions in the recommendations of scholarship recipients. This study took a sample of 1000 prospective recipients of the Sebelas Maret University PPA scholarship, as many as 907 students were classified correctly through the calculation of the algorithm *k*-NN and as many as 883 students were classified correctly through the calculation of C4.5. The results also showed that the accuracy and error of the algorithm *k*-NN was 90.7% and 9.3%, whereas the C4.5 algorithm had an accuracy and error of 88.3% and 11.7%.

Keywords: Classification, *K-Nearest Neighbor*, C4.5, Scholarship

PENDAHULUAN

Beasiswa merupakan bantuan biaya pendidikan yang ditawarkan oleh setiap perguruan tinggi. Salah satu beasiswa yang ditawarkan adalah beasiswa Peningkatan Prestasi Akademik (PPA) yang diberikan kepada mahasiswa berprestasi. Dalam pengambilan keputusan penerima beasiswa tersebut diperlukan suatu metode klasifikasi untuk meminimalkan kesalahan sasaran. Klasifikasi merupakan proses menemukan model yang menggambarkan dan membedakan kelas suatu data untuk memprediksi kelas dari suatu obyek yang lebel kelasnya belum diketahui (Han dan Kamber, 2006). Banyak teknik klasifikasi yang dapat digunakan diantaranya adalah algoritma *k*-NN dan C4.5. Penggunaan algoritma yang tepat akan menghasilkan keputusan yang akurat. Metode klasifikasi algoritma *k*-NN merupakan salah satu metode pengklasifikasian data yang memiliki konsistensi yang kuat dan efektif dalam melakukan *training* data yang besar. Algoritma C4.5 memiliki kelebihan utama yaitu dapat menghasilkan model berupa *tree* (pohon) atau aturan yang mudah diinterpretasikan, memiliki tingkat akurasi yang dapat diterima, dapat menangani variabel bertipe diskrit dan numerik. Kajian penelitian terdahulu dilakukan oleh Giat Karyono pada tahun 2016 yang mengkaji tingkat akurasi algoritma C4.5 dan *k*-NN, didapatkan hasil bahwa untuk menentukan diagnosis diabetes melitus lebih baik menggunakan algoritma *k*-NN karena tingkat akurasinya lebih tinggi dibandingkan dengan C4.5. Penelitian Heru Ismanto dan Retantyo Wardoyo pada tahun 2016 menggunakan kedua algoritma untuk menguji kompleksitas algoritma berpengaruh terhadap waktu berjalan, hasil pengujian menunjukkan bahwa waktu berjalan algoritma *k*-NN lebih stabil daripada C4.5. Banyaknya penerima beasiswa PPA yang sering kali tidak tepat sasaran diperlukan suatu teknik untuk mengklasifikasikan calon penerima beasiswa PPA sehingga lebih efektif dan dapat mengurangi

kesalahan dalam pengambilan keputusan. Algoritma C4.5 dapat membantu dalam menentukan mahasiswa yang diterima dan ditolak sebagai penerima beasiswa PPA di Universitas Sebelas Maret. Penggunaan algoritma *k*-NN dan C4.5 diharapkan dapat memberi informasi yang berguna tentang klasifikasi penerima beasiswa PPA. Tujuan dari penelitian ini adalah mengetahui hasil klasifikasi penerima beasiswa dan membandingkan akurasi dari algoritma *k*-NN dan C4.5.

Metode Klasifikasi

Proses klasifikasi didasarkan pada empat komponen berikut (Gorunescu, 2011).

1) Kelas
   Variabel dependen yang berupa kategorikal yang merepresentasikan "label" pada objek. Misalkan resiko penyakit jantung, resiko kredit macet, tinggi/rendahnya produksi.
2) Prediktor
   Variabel independen yang direpresentasikan oleh karakteristik (atribut) data. Misalkan tekanan darah, hiperkolesterol, tabungan, aset, suhu udara, luas daerah.
3) *Training Dataset*
   Satu set data yang berisi nilai dari variabel dependen dan independen yang digunakan untuk menentukan kelas yang sesuai berdasar predictor.
4) *Testing Dataset*
   Merupakan data yang akan diklasifikasikan oleh model yang telah dibuat dan akurasi klasifikasi dievaluasi.

Pohon Keputusan

Model klasifikasi yang menggunakan struktur berupa pohon dilakukan dengan mengubah bentuk data dalam suatu tabel menjadi model pohon, mengubah model pohon menjadi *rule*, dan menyederhanakan *rule*. Pohon keputusan berfungsi untuk mengeksplorasi data, menemukan hubungan tersembunyi antara variabel dependen dan independen (Han dan Kamber, 2006). Pohon keputusan mempunyai tiga simpul sebagai berikut.

1. *Root (*simpul akar): Terletak pada bagian paling atas dari pohon keputusan dan tidak memiliki *input* tetapi memiliki *output* lebih dari satu. Simpul ini berupa atribut yang memiliki pengaruh terbesar pada suatu kelas tertentu.
2. *Node* (simpul internal): Simpul ini merupakan percabangan dimana membutuhkan satu *input* dan mengeluarkan maksimal dua *output*.
3. *Leaf* (simpul daun): Simpul ini terletak pada ujung pohon dan hanya memiliki satu *input* dan tidak memiliki *output* dan menandai bahwa simpul tersebut merupakan label kelas.

Algoritma *K-Nearest Neighbor* (*k*-NN)

Algoritma *k*-NN adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data *training* yang jaraknya paling dekat dengan objek tersebut. *K*-NN adalah sebuah metode untuk mencari kasus dengan menghitung kedekatan antara kasus baru dan kasus lama. Algoritma *k*-NN adalah salah satu metode yang digunakan untuk analisis klasifikasi, namun metode *k*-NN juga digunakan untuk prediksi (Alkhatib K, 2013). Jarak antara dua titik pada data *training* dan titik pada data *testing* dapat didefinisikan dengan rumus *Euclidean*, seperti berikut.

$$d = \sqrt{\sum_{i=1}^{p} (x_{2i} - x_{1i})^2} \qquad (1)$$

dengan

d : jarak *Euclidean*

x_{2i} : nilai pada data *testing* ke - i

x_{1i} : nilai pada data *training* ke - i

p : banyaknya atribut

Normalisasi Data

Rumus *Min-Max Normalization* sebagai berikut.

$$X^* = \frac{X - \min(X)}{\max(X) - \min(X)} \tag{2}$$

dengan

X^* : data baru
X : data lama
$\min(X)$: nilai minimum dari data per kolom
$\max(X)$: nilai maksimum dari data per kolom

Algoritma C4.5

Algoritma C4.5 merupakan metode klasifikasi dan prediksi yang sangat kuat dan terkenal (Ina, 2013). Beberapa pengembangan yang dilakukan pada C4.5 adalah dapat mengatasi nilai yang hilang. Menurut Written (2007) algoritma C4.5 secara umum membangun pohon keputusan sebagai berikut.

1) Memilih variabel sebagai akar
2) Membuat cabang untuk masing-masing nilai
3) Membagi kasus dalam cabang
4) Mengulangi proses untuk masing-masing cabang sampai semua kasus pada cabang memiliki kelas yang sama

Pemilihan variabel sebagai akar, didasarkan pada nilai *gain ratio* tertinggi dari variabel-variabel yang ada. *Gain ratio* diperoleh dari perhitungan nilai *entropy, gain, dan splitinfo*. Perhitungan *entropy* digunakan rumus sebagai berikut

$$Entropy(S) = \sum_{i=1}^{n} -p_i * \log_2 p_i \tag{3}$$

dengan

S : himpunan kasus
n : jumlah partisi kasus
p_i : proporsi dari S_i

Nilai *entropy* yang telah dihitung digunakan dalam perhitungan nilai *gain* sebagai berikut.

$$Gain(S, A) = Entropy - \sum_{i=1}^{n} \frac{|S_i|}{S} * EntropyS_i \tag{4}$$

dengan

S : himpunan kasus
A : atribut
n : jumlah partisi atribut
$|S_i|$: jumlah kasus pada partisi ke-i

Langkah berikutnya menghitung *SplitInfo* dengan rumus sebagai berikut.

$$SplitInfo(S, A) = -\sum_{i=1}^{n} \frac{S_i}{S} \log_2 \frac{S_i}{S} \tag{5}$$

Gain Ratio dihitung dengan rumus berikut.

$$GainRatio(S, A) = \frac{Gain(S, A)}{SplitInfo(S, A)} \tag{6}$$

Confusion Matrix

Dataset terdiri dari dua kelas, kelas yang satu dianggap positif dan yang lain negatif (Bramer, 2007) berupa tabel matriks yang dapat dilihat pada **Tabel 1**.

Tabel 1. ***Confusion Matrix***

Klasifikasi yang benar	Diklasifikasikan sebagai	
	+	-
+	*True positive*	*False negative*
-	*False positive*	*True negative*

Akurasi

Tingkat akurasi sebuah klasifikasi adalah rasio perbandingan jumlah data *testing* yang dapat diklasifikasikan dengan benar dengan jumlah seluruh data *testing* (Bahri, 2012). Tingkat akurasi dapat dinyatakan dalam persamaan berikut.

$$Akurasi = \frac{banyaknya\ prediksi\ yang\ benar}{total\ banyak\ prediksi} * 100\% \quad (7)$$

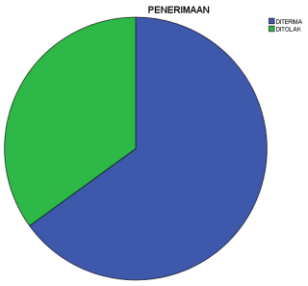
METODE PENELITIAN

Penelitian ini menggunakan data pendaftar beasiswa PPA 2018 yang diperoleh dari pihak Biro Kemahasiswaan UNS. Data pendaftar beasiswa PPA ini meliputi sepuluh fakultas di UNS. Variabel dependen yang digunakan dalam penelitian ini adalah diterima/ tidak mahasiswa tersebut. Variabel independen yang digunakan adalah IPK, semester, jumlah orang tua bekerja, penghasilan orang tua, jumlah tanggungan, dan prestasi.

Metode penelitian yang digunakan berdasarkan kerangka kerja yang terdiri dari langkah-langkah sebagai berikut.

A. Mendeskripsikan data pendaftar beasiswa PPA UNS

Dalam menentukan penerima beasiswa PPA menggunakan algoritma *k*-NN dan C4.5, data yang digunakan dan dianalisa adalah data calon penerima beasiswa PPA UNS tahun 2018. Variabel independen yang digunakan yaitu IPK, semester, jumlah orang tua bekerja, penghasilan orang tua, jumlah tanggungan, dan prestasi. Variabel dependen yang digunakan adalah penerimaan beasiswa dengan output diterima atau ditolak. Dari 1000 data calon penerima beasiswa, sebanyak 65% diterima dan 35% ditolak yang disajikan pada Gambar 1. berikut



Gambar 1. Presentase penerimaan beasiswa PPA

B. Menyusun dan mengategorikan data pada variabel dependen dan independen

Variabel dan nilai variabel dapat dilihat pada Tabel 2.

Tabel 2. Kategori Variabel

NO	VARIABEL	NILAI VARIABEL
1	Indeks Prestasi Kumulatif (IPK)	1. 3.00 – 3.20 2. 3.21 – 3.40 3. 3.41 – 3.60 4. 3.61 – 4.00
2	Semester	1. 1 - 2 2. 3 - 4 3. 5 - 6 4. 7 - 8
3	Prestasi	1. Tidak ada 2. Wilayah/Propinsi 3. Nasional 4. Internasional
4	Jumlah Orang tua bekerja	1. Ayah/Ibu 2. Ayah dan Ibu
5	Penghasilan	1. ≤ 1500000 2. $1500000 < X < 2500000$ 3. $2500000 \leq X < 3500000$ 4. ≥ 3500000
6	Jumlah Tanggungan	1. 1 2. 2 - 3 3. > 3
7	Penerimaan	Diterima Ditolak

C. Melakukan klasifikasi menggunakan Algoritma *k*-NN
   a. Normalisasi Data
   b. Menentukan parameter *k* dengan mengambil nilai *k*<10. Misalkan memilih *k* =1
   c. Menghitung jarak *euclid*
   d. Mengurutkan jarak *euclid* dari yang terkecil dan mengumpulkan kategori dependen
   e. Menggunakan kategori mayoritas untuk mendapatkan hasil klasifikasi

D. Melakukan klasifikasi menggunakan algoritma C4.5
   Gain Ratio tertinggi dipilih menjadi *root node*, selanjutnya mengulangi langkah sama sampai semua *record* terpartisi.

HASIL DAN PEMBAHASAN

1. Hasil algoritma *k*-NN

Berdasarkan hasil pencarian *k*-optimal didapatkan *k*=1. Selanjutnya adalah menghitung jarak *euclid* data *testing* terhadap data *training* dan dihasilkan pada Tabel 3 sebagai berikut.

Tabel 3. Perhitungan jarak *euclid*

N	Jarak *euclid*	Penerimaan
1	5,174725	DITERIMA
2	5,577734	DITERIMA
3	5,656854	DITERIMA
4	3,901567	DITOLAK
5	3,901567	DITERIMA
.....	.....	.....
1000	4,818944	DITERIMA

Hasil perhitungan jarak euclid dengan data training sejumlah 1000 data diurutkan mulai dari terkecil hingga terbesar . Selanjutnya dilihat mayoritas klasifikasi yang muncul dari perhitungan jarak pertama, selanjutnya lakukan hal serupa untuk seluruh data *testing*. Dari perhitungan tersebut, didapatkan *confusion matrix* untuk algoritma *k*-NN pada Tabel 4 sebagai berikut.

Tabel 4. ***Confusion Matrix k*-NN**

	Pred. Diterima	*Pred.* Ditolak
True Diterima	740	12
True Ditolak	81	167

Data yang diprediksi dengan benar melalui algoritma *k*-NN sejumlah 907, sebanyak 740 mahasiswa diterima beasiswa PPA dan 167 mahasiswa ditolak. Sebanyak 12 mahasiswa diterima tetapi diklasifikasikan ditolak dan sebanyak 81 mahasiswa ditolak tetapi diklasifikasikan diterima.

Akurasi yang dihasilkan sebagai berikut.

$$Akurasi = \frac{740+167}{1000} * 100\% = 90.7\%$$

2. Hasil algoritma C4.5

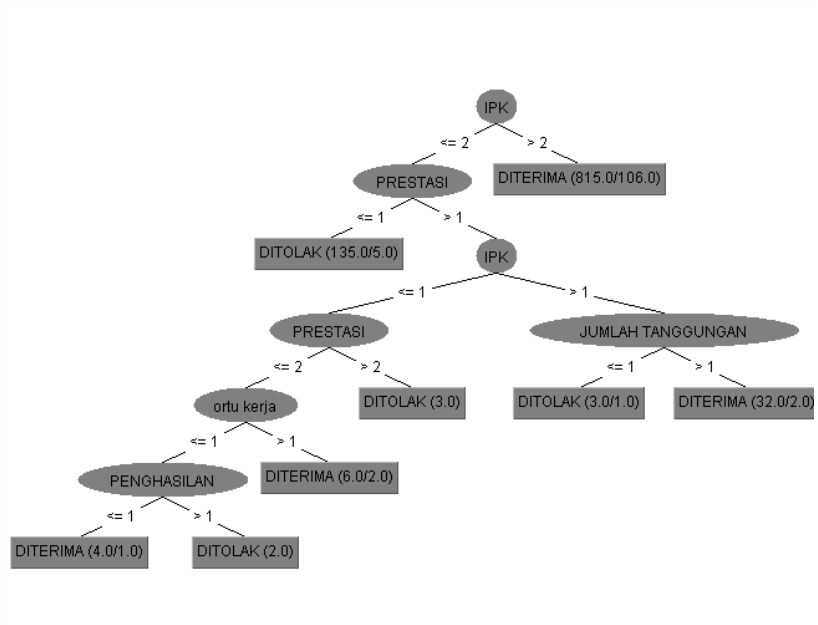
   Didapatkan hasil perhitungan *Gain Ratio node* 1pada Tabel 4 sebagai berikut.

Tabel 5. *Gain Ratio*

Variabel	Kategori	*Entropy*	*Information Gain*	*SplitInfo*	***Gain Ratio***
Penerimaan		0,808093			
IPK	1	0,563103	0,226802	1,663272	**0,136359**
	2	0,845351			
	3	0,693777			
	4	0,404448			
Semester	1	0,908924	0,102677	1,790118	0,057358
	2	0,810057			
	3	0,700967			
	4	0,680077			
Prestasi	1	0,873794	0,039694	0,984871	0,040304
	2	0,329846			
	3	0,468996			
	4	0,266765			
Jumlah orang tua bekerja	1	0,775029	0,002447	1,912749	0,002558
	2	0,856027			
Penghasilan	1	0,650022	0,033046	0,956617	0,017277
	2	0,628676			
	3	0,784194			
	4	0,945465			
Jumlah tanggungan	1	0,797327	0,000833	1,19691	0,000696
	2	0,840189			
	3	0,790687			

Dari perhitungan *Gain Ratio* diatas, variabel yang menjadi *node* akar adalah IPK dengan *Gain Ratio* tertinggi yaitu 0.136359. Langkah berikutnya, menghitung *gain ratio* untuk mendapatkan

node selanjutnya hingga semua *record* terpartisi. Algoritma C4.5 menghasilkan suatu keputusan berbentuk pohon yang dihasilkan dari software WEKA sebagai berikut.



Gambar 2. Pohon keputusan penerima beasiswa

Aturan Klasifikasi :
1. Jika IPK lebih besar dari 3.40 maka diterima
2. Jika IPK kurang dari sama dengan 3.40, prestasi tidak ada maka ditolak
3. Jika IPK lebih dari 3.20, prestasi ada, dan jumlah tanggungan lebih dari satu maka diterima
4. Jika IPK lebih dari 3.20, prestasi ada, dan jumlah tanggungan satu maka ditolak
5. Jika IPK kurang dari sama dengan 3.20, berprestasi di nasional/internasional maka ditolak
6. Jika IPK kurang dari sama dengan 3.20, tidak ada prestasi atau berprestasi di wilayah/propinsi, orang tua yang bekerja satu, penghasilan kurang dari sama dengan 1,5 juta maka diterima
7. Jika IPK kurang dari sama dengan 3.20, tidak ada prestasi atau berprestasi di wilayah/propinsi, orang tua yang bekerja satu, penghasilan lebih dari 1,5 juta maka ditolak
8. Jika IPK kurang dari sama dengan 3.20, tidak ada prestasi atau berprestasi di wilayah/propinsi, orang tua yang bekerja satu maka diterima

Pohon keputusan yang dihasilkan menunjukkan bahwa *node* akar adalah IPK. Hasil ini sebanding dengan perhitungan melalui *Gain Ratio* tertinggi sebesar 0.136359. Mahasiswa dengan IPK lebih dari 3.40 merupakan mahasiswa yang diterima beasiswa PPA, sedangkan mahasiswa dengan nilai IPK kurang dari sama dengan 3.40 harus memiliki kriteria lanjutan berdasar variabel-variabel yang berpengaruh seperti prestasi, jumlah tanggungan, penghasilan, dan jumlah orang tua bekerja. Dari pohon keputusan diatas dapat diketahui *confusion matrix* algoritma C4.5 pada Tabel 6 sebagai berikut.

Tabel 6. *Confusion Matrix* C4.5

	Pred. Diterima	*Pred.* Ditolak
True Diterima	746	6
True Ditolak	111	137

Data yang diprediksi dengan benar melalui algoritma C4.5 sejumlah 883, sebanyak 746 mahasiswa diterima beasiswa PPA dan 137 mahasiswa ditolak. Sebanyak 6 mahasiswa diterima tetapi diklasifikasikan ditolak dan sebanyak 111 mahasiswa ditolak tetapi diklasifikasikan diterima.
Akurasi yang dihasilkan sebagai berikut.

$$Akurasi = \frac{746+137}{1000} * 100\% = 88.3\%$$

Perbandingan hasil akurasi algoritma *k*-NN dan C4.5 pada Tabel 7.

Tabel 7. Perbandingan Akurasi *k*-NN dan C4.5

Algoritma	Akurasi
k-NN	90.7%
C4.5	88.3%

Berdasarkan akurasi dari algoritma *k*-NN dan C4.5, data calon penerima beasiswa PPA lebih baik menggunakan algoritma *k*-NN karena memiliki akurasi lebih tinggi dari algoritma C4.5. Hal ini dipengaruhi oleh pemilihan k-optimal pada algoritma *k*-NN sehingga memiliki tingkat akurasi yang lebih tinggi.

SIMPULAN DAN SARAN

Berdasarkan penelitian yang dilakukan, dapat disimpulkan bahwa dalam merekomendasikan penerima beasiswa PPA dapat dilakukan menggunakan algoritma *k*-NN dan C4.5 dengan akurasi masing-masing sebesar 90.7% dan 88.3%. Hasil penelitian menunjukkan bahwa variabel memiliki pengaruh adalah IPK, prestasi, penghasilan, jumlah orang tua bekerja, dan jumlah tanggungan. Variabel yang paling berpengaruh adalah variabel IPK. Dalam mengukur kedua fungsi algoritma tersebut menggunakan *confusion matrix* dengan hasil bahwa algoritma *k*-NN memiliki tingkat akurasi yang lebih tinggi dibandingkan algoritma C4.5. Dari 1000 data calon penerima beasiswa PPA, sebanyak 907 mahasiswa diklasifikasikan secara benar melalui perhitungan algoritma *k*-NN dan sebanyak 883 mahasiswa diklasifikasikan secara benar melalui perhitungan C4.5. Pada penelitian selanjutnya dapat menambah variabel dan membandingkan dengan algoritma lainnya seperti CHAID, C5.0, *Naive Bayes.*

DAFTAR PUSTAKA

Alkhatib K, Najadat H, Hmeidi I, Shatnawi MKA. 2013. Stock Price Prediction Using k-Nearest Neighbor (k-NN) Algorithm. *International Journal of Business, Humanities and Technology*. Vol. 3, No. 3, 32-44.

Bahri R, Sofian. 2012. Perbandingan Algoritma Template Matching Dan Feature Extraction Pada Optical Character Recognition. *Jurnal Computer Dan Informatika.* Universitas Computer Indonesia, Edisi. I, Vol.1.

Bode, Andi. 2017. K-Nearest Neighbor Dengan Feature Selection Menggunakan Backward Elimination Untuk Prediksi Harga Komoditi Kopi Arabika. *Jurnal Ilmiah.*Volume 9 Nomor 2 Agustus 2017.

Bramer, M. 2007. *Principles of Data Mining*. Springer, London.

Gorunescu, F. 2011. *Data Mining Concepts, Models and Techniques*, Verlah Berlon Heidelberg: Spinger.

Han, J. And M. Kamber. 2006. *Data mining Concept and Tehniques*. San Fransisco: Morgan Kauffman.

Ismanto. H, Wardoyo. R. 2016. Analysis Of C4.5 And K-Nearest Neighbor (KNN) Method On Algorithm Of Clustering For Deciding Mainstay Area. *IOSR Journal of Computer Engineering (IOSR-JCE)* e-ISSN: 2278-0661,p-ISSN: 2278-8727, Vol.18, Issue 2, Ver. IV (Mar-Apr. 2016), PP 86-92.

Karyono, Giat. 2016. Analisis Teknik Data Mining Algoritma C4.5 Dan K-Nearest Neighbor Untuk Mendiagnosa Penyakit Diabetes Mellitus. STMIK – Politeknik PalComTech, 12 Mei 2016.

Larose, D.T. 2005. *Discovering Knowledge in Data*. New Jersey, John Willey & Sons, Inc.

Lestari, Mei. 2014. Penerapan Algoritma Klasifikasi Nearest Neighbor (K-NN) untuk mendeteksi penyakit jantung. Faktor Exacta Vol. 7, No. 4, 366-371, ISSN:1979-276X, 2014.

Sumarlin. 2015. Implementasi Algoritma K-Nearest Neighbor Sebagai Pendukung Keputusan Klasifikasi Penerima Beasiswa PPA dan BBM, *Jurnal Sistem Informasi Bisnis 01*.
Written, Frank. 2007. *Data Mining Practical Machine Learning tools and Techniques.* USA.